

A HIERARCHICAL BAYES APPROACH FOR ESTIMATING SMALL AREA PROPORTIONS

P.J. Farrell¹

ABSTRACT

The importance of small area estimation as a facet of survey sampling cannot be over-emphasized. Of late, there has been an increasing demand for small area statistics in both the public and private sectors. It is widely recognized that direct survey estimators for small areas are likely to be unstable due to the small (or nonexistent) sample sizes taken from these areas. This makes it necessary to "borrow strength" from related areas to obtain estimators which are less variable.

In this study, a hierarchical Bayes model-based methodology for the estimation of small area proportions is proposed, implemented, and evaluated. The basic idea consists of incorporating into a logistic regression model containing predictor variables, random effects which reflect the structure of the sample design. Estimates for the model parameters are obtained using the griddy-Gibbs sampler developed by Ritter and Tanner (1992). Two hierarchical Bayes estimators are derived which use these model estimates to determine point and interval estimates for small area proportions. One requires predictor variable data for all individuals within a local area, the other is based on local area summary statistics for these variables. The proposed estimators are applied to data from a United States Census to predict local labour force participation rates.

RÉSUMÉ

L'importance de l'estimation sur des petites régions comme une facette de l'enquête par échantillonnage, ne peut être excessivement soulignée. Récemment, une demande croissante est apparue pour des statistiques sur des petites régions dans les domaines public et privé. On reconnaît largement que les estimateurs directs à partir d'enquêtes pour des petites régions sont probablement instables, à cause des petites (ou inexistantes) tailles des échantillons pris de ces régions. Ceci rend nécessaire un "emprunt" à partir de régions apparentées pour obtenir des estimateurs qui soient moins variables.

Dans cette étude, une méthodologie basée sur un modèle de Bayes hiérarchique pour l'estimation de proportions sur des petites régions est proposée et évaluée. L'idée de base consiste à incorporer des effets aléatoires qui reflètent la structure du modèle de l'échantillon à l'intérieur d'un modèle de régression logistique contenant des variables de prédiction. Les estimations des paramètres modèles sont obtenues en utilisant la sélection "griddy-Gibbs" développée par Ritter et Tanner (1992). On calcule deux estimateurs de Bayes hiérarchiques qui utilisent ces estimations de modèle pour déterminer les estimations ponctuelles et par intervalle pour les proportions des petites régions. L'un nécessite des données de variable prédictive pour tous les individus d'un territoire local, l'autre est basé sur les statistiques sommaires des territoires locaux pour ces variables. Les estimateurs proposés sont appliqués aux données d'un Recensement des État-Unis pour prédire des taux de participation de main d'oeuvre locale.

1. INTRODUCTION

The terms "small area" and "local area" are commonly used to denote a small geographic area, such as a county, a municipality, or a census division. Of late, there has been an increasing demand for small area statistics in both the public and private sectors.

Unfortunately, a number of the estimators proposed for small area parameters are deficient in certain aspects. The usual direct survey estimators for a small area, based on data only from the sampled units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area. This deficiency has led to the development of model-based estimation approaches which "borrow strength" from

related local areas to provide estimators which are more accurate. Ghosh and Rao (1994) reviewed and compared the available techniques for small area estimation using simulated wage and income data. They showed that empirical and hierarchical Bayes techniques, for most purposes, seem to have a distinct advantage over other methods. Similar conclusions were reached by Farrell, MacGibbon, and Tomberlin (1994a) in a study which compared an empirical Bayes estimator with an unbiased direct survey estimator and a synthetic estimator.

Several authors have considered the problem of estimating small area rates and binomial parameters using empirical and hierarchical Bayes approaches. Dempster and Tomberlin (1980) proposed an empirical Bayes method for estimating census undercount for local areas

¹ Patrick J. Farrell, Assistant Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1.

based on logistic regression models containing fixed and random effects. This proposal was further developed by MacGibbon and Tomberlin (1989) and Farrell, MacGibbon, and Tomberlin (1994a); the objective of these authors was the estimation of small area proportions. Others have considered hierarchical Bayes approaches. Stroud (1991) studied hierarchical Bayes models for univariate natural exponential families with quadratic variance functions, while Malec, Sedransk, and Tompkins (1993) used a fully Bayes approach based on a logistic regression model to estimate proportions using data from the National Health Interview Survey.

In this study, a hierarchical Bayes model-based methodology for the estimation of small area proportions is proposed, implemented, and evaluated. The basic idea consists of incorporating into a logistic regression model containing predictor variables, random effects which reflect the structure of the sample design. Estimates for the model parameters are obtained using the griddy-Gibbs sampler developed by Ritter and Tanner (1992). Two hierarchical Bayes estimators are derived which use these model estimates to determine point and interval estimates for small area proportions. One requires predictor variable data for all individuals within a local area, the other is based on local area summary statistics for these variables. As an illustration of the proposed methodology, these two estimators are applied to data from a United States Census to predict local labour force participation rates for females.

The proposed estimation procedures are described in Section 2. A data example is presented in Section 3, while the conclusions and discussion are given in Section 4.

2. ESTIMATION PROCEDURES

The objective of this study is the development of point and interval estimates for small area proportions using a two stage sample design. Let p_i represent the proportion of individuals in the i -th local area which possess a characteristic of interest. Then

$$p_i = \frac{\sum_j y_{ij}}{N_i}, \quad (2.1)$$

where N_i is the population size of local area i , and y_{ij} takes on a value of zero or one, depending upon whether or not the j -th individual within the i -th local area possesses the characteristic of interest.

We wish to estimate the parameters p_i . A predictive model-based approach proposed by Royall (1970) is used to specify an estimator. Under this approach, the estimator of p_i in (2.1) is:

$$\hat{p}_i = \frac{\sum_{j \in S} y_{ij} + \sum_{j \in S'} \hat{y}_{ij}}{N_i}, \quad \text{where } \sum_{j \in S'} \hat{y}_{ij} = \sum_{j \in S'} \hat{\pi}_{ij}, \quad (2.2)$$

the sum over $j \in S$ of y_{ij} is the sum of the values of the outcome variable for sampled individuals from the i -th local area, and the sum over $j \in S'$ of $\hat{\pi}_{ij}$ is the sum of the estimated probabilities for nonsampled individuals in the i -th local area.

To obtain values for $\hat{\pi}_{ij}$, an explicitly model-based approach is employed. Under this approach, a model which describes the probabilities π_{ij} associated with individuals in the population is as follows:

$$y_{ij} | \pi_{ij} \sim \text{i.i.d. Bernoulli}(\pi_{ij}), \quad (2.3)$$

$$\text{logit}(\pi_{ij}) = \underline{X}_{ij}^T \underline{\beta} + \delta_i.$$

The vector \underline{X}_{ij} represents a vector of fixed effects predictor variables which is augmented by the constant one, while $\underline{\beta}$ is a vector of fixed effect logistic regression parameters which contains the constant term β_0 . The quantity δ_i is a random effect associated with the i -th local area. The δ_i are sampled from some prior probability distribution.

Here we are interested in obtaining point and interval estimates for local area proportions, p_i . We require estimates of π_{ij} , where

$$\pi_{ij} = [1 + \exp\{-\underline{X}_{ij}^T \underline{\beta} + \delta_i\}]^{-1}. \quad (2.4)$$

Once hierarchical Bayes estimates $\hat{\underline{\beta}}$ and $\hat{\delta}_i$ have been determined, π_{ij} is estimated by

$$\hat{\pi}_{ij} = [1 + \exp\{-\underline{X}_{ij}^T \hat{\underline{\beta}} + \hat{\delta}_i\}]^{-1}. \quad (2.5)$$

2.1 Parameter estimates

A joint multivariate normal prior distribution is assumed for the random effects. Thus, the model in (2.3) becomes:

$$\text{logit}(\pi_{ij}) = \underline{X}_{ij}^T \underline{\beta} + \delta_i,$$

$$\delta_i \sim \text{i.i.d. Normal}(0, \tau^2), \quad (2.6)$$

$$\tau^2 \sim \text{Inverse } \Gamma(a, b).$$

More specifically, the random effects are assumed to follow normal distributions, each with a mean of zero, and the same unknown variance τ^2 . Since a hierarchical Bayes approach is employed to estimate the parameters in the above model, it is also necessary to specify a distribution for the parameter τ^2 of the prior distribution.

Here, a diffuse version of an inverse gamma distribution is assumed for τ^2 , where the parameters of the distribution are set to zero.

To develop estimates for the model in (2.6), we begin by considering the joint distribution of the data and parameters. Let n be the number of sampled local areas. If a flat prior is placed upon the fixed effects parameters, then the joint distribution of the data and the parameters is given by

$$f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2) \propto \prod_{ij} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1 - y_{ij}} \frac{1}{\tau^n} \times \exp(-\frac{1}{2} \sum_i \delta_i^2 / \tau^2) \frac{b^a \exp(-b / \tau^2)}{\tau^{2(a+1)} \Gamma(a)}, \quad (2.7)$$

where $\underline{\delta}$ and \underline{y} are vectors containing δ_i and the data y_{ij} , respectively. This joint distribution can be employed to determine posterior distributions for the components of $\underline{\beta}$ and $\underline{\delta}$, as well as for τ^2 :

$$f(\beta_u | \underline{y}, \beta_0, \beta_1, \dots, \beta_{u-1}, \beta_{u+1}, \dots, \beta_m, \underline{\delta}, \tau^2) = \frac{f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2)}{\int f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2) d\beta_u}, \quad (2.8)$$

$$f(\delta_i | \underline{y}, \underline{\beta}, \delta_1, \dots, \delta_{i-1}, \delta_{i+1}, \dots, \delta_n, \tau^2) = \frac{f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2)}{\int f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2) d\delta_i}, \quad (2.9)$$

$$f(\tau^2 | \underline{y}, \underline{\beta}, \underline{\delta}) = \frac{f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2)}{\int f(\underline{y}, \underline{\beta}, \underline{\delta}, \tau^2) d\tau^2}, \quad (2.10)$$

where u is the number of predictor variables, and $u = 0, 1, \dots, m$.

It is not feasible to obtain closed form expressions for the posteriors given in (2.8) through (2.10) due to the intractable integrations required to evaluate the denominators. However, these equations can be used to define the following series of expressions that the posteriors will be proportional to

$$f(\beta_u | \underline{y}, \beta_0, \beta_1, \dots, \beta_{u-1}, \beta_{u+1}, \dots, \beta_m, \underline{\delta}, \tau^2) \propto \prod_{ij} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1 - y_{ij}}, \quad (2.11)$$

$$f(\delta_i | \underline{y}, \underline{\beta}, \delta_1, \dots, \delta_{i-1}, \delta_{i+1}, \dots, \delta_n, \tau^2) \propto \prod_j \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1 - y_{ij}} \exp(-\frac{1}{2} \delta_i^2 / \tau^2), \quad (2.12)$$

$$f(\tau^2 | \underline{y}, \underline{\beta}, \underline{\delta}) \propto \frac{1}{\tau^n} \exp(-\frac{1}{2} \sum_i \delta_i^2 / \tau^2) \times \frac{\exp(-b / \tau^2)}{\tau^{2(a+1)}} = \frac{1}{\tau^{n+2}} \exp(-\frac{1}{2} \sum_i \delta_i^2 / \tau^2), \quad (2.13)$$

provided that a diffuse version of the inverse gamma distribution is placed upon

τ^2 , where a and b are both set to zero.

To determine estimates for $\underline{\beta}$, $\underline{\delta}$, and τ^2 , the griddy-Gibbs sampler proposed by Ritter and Tanner (1992) is used in conjunction with the expressions in (2.11) through (2.13). Provided that P different paths of T iterations each of the griddy-Gibbs sampler are performed, P sets of model estimates will be available upon convergence of the algorithm.

2.2 Estimates of small area proportions

Once the hierarchical Bayes model estimates have been obtained, (2.2) is used to obtain hierarchical Bayes point estimates for small area proportions. Since P sets of these estimates have been obtained using the griddy-Gibbs sampler, it is possible to use (2.2) to compute P values for the estimator \hat{p}_i , which are then treated as an empirical distribution. The 50th percentile of this distribution can be taken as the point estimate of the proportion of the i -th local area. In addition, if a $100(1 - \alpha)\%$ interval estimate, say 95%, is also required, the range bounded by the 2.5th and 97.5th percentiles of the distribution can be taken as one possible interval estimate.

In the usual case when micro-data concerning predictor variables for all individuals in a local area are not available, \hat{p}_i in (2.2) cannot be determined. Using local area summary statistics for predictor variables, Farrell, MacGibbon, and Tomberlin (1994b) proposed a model-based procedure for estimating binomial small area parameters which required only local area summary statistics for predictor variables.

To obtain an estimator, \bar{p}_i , for the i -th local area using only local area summary statistics for predictor variables, a second-order multivariate Taylor series expansion of $\sum \hat{\pi}_{ij}$ in (2.2) is taken about $\bar{X}_{-i,S'}$, the mean vector of the fixed effects predictor variables for nonsampled individuals in local area i .

The details of the second order Taylor series expansion are given in Farrell, MacGibbon, and Tomberlin (1994b), and are not repeated here. However, in addition to $\bar{X}_{-i,S'}$, the finite population covariance matrix, $V_{-i,S'}$, of the auxiliary variables for nonsampled individuals in local area i is also required. The Taylor series expansion yields the following approximation for the estimator \hat{p}_i in (2.2):

$$\bar{p}_i = N_i^{-1} \sum_{j \in S} y_{ij} + (N_i - n_i) \times \left\{ \bar{p}_i + \frac{1}{2} [\bar{p}_i(1-\bar{p}_i)^2 - \bar{p}_i^2(1-\bar{p}_i)] \hat{\beta}^T \mathbf{V}_{i,S} \hat{\beta} \right\} \quad (2.14)$$

where

$$\bar{p}_i = [1 + \exp \{ -(\bar{\mathbf{X}}_{i,S}^T \hat{\beta} + \hat{\delta}_i) \}]^{-1}. \quad (2.15)$$

The same approach used to derive point and interval estimates for local area proportions using \hat{p}_i can also be employed with \bar{p}_i .

3. A DATA EXAMPLE

The proposed methods are now used to estimate local labour force characteristics using data from a 1% sample of the 1950 United States Census (United States Bureau of the Census 1984). Here, attention centers on the estimation of female labour force participation rates for local areas, where local areas are more or less confined to states.

The data for estimating local area female labour force participation rates were obtained from the 1% sample using a two stage sample design. In the first stage, twenty local areas were selected without replacement

using probabilities proportional to size (PPS). Then, 50 individuals were randomly selected from each chosen local area, bringing the total sample size to 1,000. Two hundred samples were drawn in this fashion, making it possible to study the properties of the two estimators \hat{p}_i and \bar{p}_i over repeated realizations of the sample design. However, since inference is to be based conditionally on a particular set of local areas being sampled, resampling was not performed at the local area selection stage.

The model employed for the data example is given by (2.6). The predictor variables, selected using a stepwise logistic regression procedure on a simple random sample of 2,000 females, were age, marital status, and whether a female had children.

For each of the two hundred replicates, one hundred paths of the griddy-Gibbs sampler were performed. Each path consisted of a series of two hundred iterations of the sampler.

For both \hat{p}_i and \bar{p}_i , average estimated rates (over all 200 replicates) for each of the twenty sampled local areas are presented in Figure 1, arranged in ascending order according to the population rates. The population rates are also plotted. In order to evaluate the design bias of \hat{p}_i and \bar{p}_i , over all sampled local areas, the mean absolute difference between the small area proportions and the

FIGURE 1
Average Point Estimates for the Proportions of Sampled Local Areas

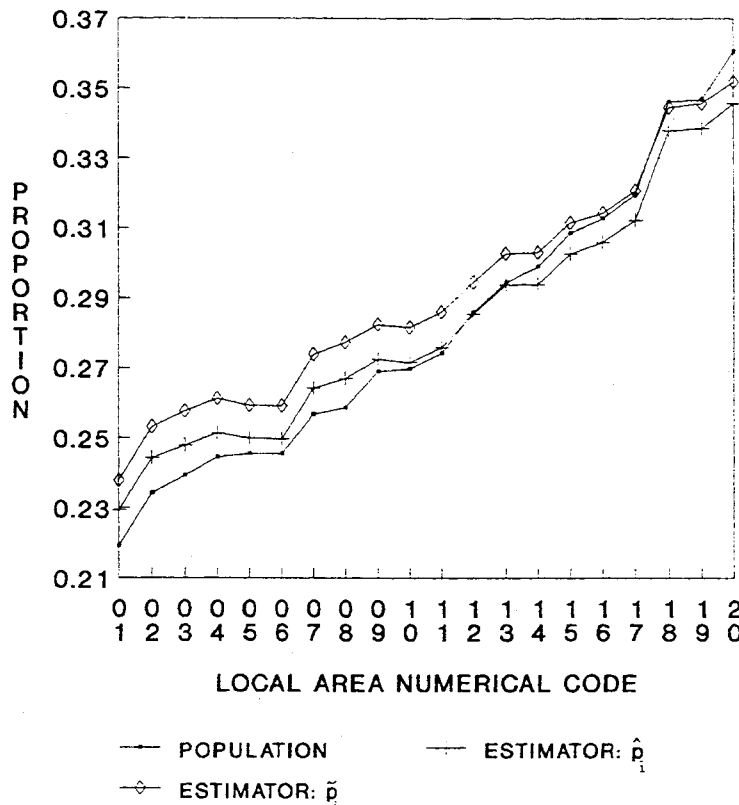
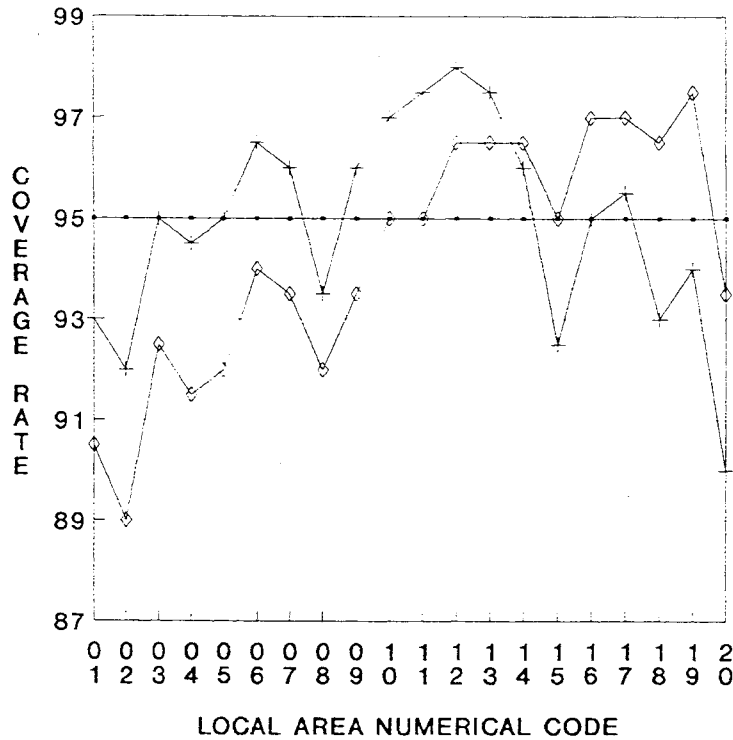


FIGURE 2
Coverage Rates for Sampled Local Areas



+ ESTIMATOR: \hat{p}_i
◇ ESTIMATOR: \bar{p}_i

average estimated rates was determined for each estimator. The mean absolute differences obtained for rates based on \hat{p}_i and \bar{p}_i were 0.0062 and 0.0105, respectively. Thus, both estimators perform quite well in terms of design bias.

The coverage properties of 95% interval estimates obtained using \hat{p}_i and \bar{p}_i for the two hundred replicates were also studied. The coverage rate for each sampled local area is presented in Figure 2. Across all sampled local areas, the average coverage rates for the 95% interval estimates based on \hat{p}_i and \bar{p}_i were 94.88% and 94.23%, respectively. Thus, the average coverage rates associated with both estimators are both extremely close to the 95% nominal rate.

Thirty-two of the fifty-two local areas were not sampled. The two hierarchical Bayes estimators were also used to predict labour force participation rates for females in these nonsampled local areas. As before when sampled local areas were considered, the results obtained using the hierarchical Bayes estimator \hat{p}_i are slightly, but not dramatically better. The mean absolute difference between the small area proportions and the average estimated rates for \hat{p}_i was 0.0298, while the analogous difference for \bar{p}_i was 0.0350. In addition, over the thirty-two nonsampled local areas, the average coverage rates

for 95% interval estimates based on \hat{p}_i and \bar{p}_i were 94.72% and 93.56%, respectively.

4. CONCLUSION

Two hierarchical Bayes model-based estimators have been derived and applied to data from the 1950 United States Census to predict local labour force participation rates for females. Both estimators performed well in terms of design bias and coverage rates. Results obtained for the estimator based on summary statistics for predictor variables were gratifyingly close to those obtained for the estimator requiring micro-data. Since information on predictor variables must often be based on Census data, such micro-data would not be available for most surveys. The estimator based on summary statistics should prove useful in these situations.

ACKNOWLEDGEMENT

The author would like to acknowledge the financial support of NSERC of Canada.

REFERENCES

- Dempster, A.P., and Tomberlin, T.J. (1980). "The Analysis of Census Undercount from a Postenumeration Survey", *Proceedings of the Conference on Census Undercount*, Arlington, VA, 88-94.
- Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1994a). "Empirical Bayes Estimators of Small Area Proportions in Multistage Designs", Currently under review for publication in *Statistica Sinica*.
- Farrell, P.J., MacGibbon, B., and Tomberlin, T.J. (1994b). "Empirical Bayes Small Area Estimation Using Logistic Regression Models and Summary Statistics", Currently under review for publication in the *Journal of Business and Economic Statistics*.
- Ghosh, M., and Rao, J.N.K. (1994). "Small Area Estimation: An Appraisal", *Statistical Science*, 9, 552-93.
- MacGibbon, B., and Tomberlin, T.J. (1989). "Small Area Estimates of Proportions via Empirical Bayes Techniques", *Survey Methodology*, 15, 237-252.
- Malec, D., Sedransk, J., and Tompkins, L. (1993). "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interview Survey", in *Case Studies in Bayesian Statistics*, Edited by Constantine Gatsonis, James S. Hodges, Robert Kasf, and Nozer D. Singpurwalla, New York: Springer-Verlag.
- Ritter, C., and Tanner, M.A. (1992). "Facilitating The Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler", *Journal of the American Statistical Association*, 87, 861-868.
- Royall, R.M. (1970). "On Finite Population Sampling Theory Under Certain Linear Regression Models", *Biometrika*, 57, 377-387.
- Stroud, T.W.F. (1991). "Hierarchical Bayes Predictive Means and Variances with Application to Sample Survey Inference", *Communications in Statistics, Theory and Methods*, 20, 13-36.
- United States Bureau of the Census (1984). "Census of the Population, 1950: Public Use Microdata Sample Technical Documentation", edited by J.G. Keane, Washington, D.C.